Reference: *Biol. Bull.* **195:** 251–254. (December, 1998)

Consensus Translational Initiation Sites of Marine Invertebrate Phyla

RAJ V. MANKAD, ALEXANDER A. GIMELBRANT, AND TIMOTHY S. MCCLINTOCK*

Department of Physiology, University of Kentucky, 800 Rose St., Lexington, Kentucky 40536-0298

The efficiency of translational initiation depends upon the sequence context surrounding the AUG codon (1, 2, 3). A purine at position −3 contributes critically to context, but other neighboring nucleotides are also important. Nucleotide frequencies at these neighboring positions vary among distant taxa (4, 5). We have analyzed the translational initiation sites of cnidarian, echinoderm, molluscan, annelid, and crustacean sequences in nucleotide sequence databases. These taxa conform to the pattern of a strong preference for a purine at −3, but the frequencies of nucleotides at neighboring positions are characteristic for each taxon. The consensus translational initiation sequences of the marine invertebrate taxa are also different from those of vertebrates and single-celled eukaryotes. These consensus sequences are useful guides for predicting translational initiation sites in cDNA clones.

The initiation of translation in eukaryotes requires the function of several complexes of proteins (6, 7, 8, 9, 10). In cap-dependent translation, a ternary complex of eukaryotic initiation factor 2B (eIF2B), GTP, and Met-tRNA interacts with a 40 S ribosomal subunit complex containing eIF1A and eIF3 to form the 43 S preinitiation complex. The 43 S complex is recruited to the cap at the 5′ end of the mRNA by eIF4F (composed of eIF4E, eIF4G, and eIF4A). The 43 S complex then scans the mRNA in the 3′ direction. The melting of RNA secondary structure, which can interfere with binding of the 43 S complex to the RNA and with scanning, is accomplished by the helicase activity of eIF4A and eIF4B. When an AUG is encountered, the 43 S complex pauses or slows. During the pause, the hydrolysis of GTP by eIF2 is associated with the release of the initiation factors and the binding of the 60 S ribosomal subunit to the 40 S subunit to form the 80 S ribosome, and translation is initiated. Because the 43 S complex scans the mRNA from the 5′ end to the 3′, the most 5′ AUG is often the site of initiation of translation. However, mRNAs in which downstream AUGs are translational initiation sites are common. In cases where an upstream AUG codon is followed by a short open reading frame, the 40 S subunit can reinitiate scanning and initiate translation from a downstream AUG (11). For some mRNAs, a cap-independent mechanism of translation is used, and upstream AUGs can be bypassed completely by the initiation complex (10). In addition, some of the initiation factors are subject to physiological regulation by signaling pathways, supporting the idea that the initiation of translation is a significant component of the regulation of gene expression in cells (6, 12). For further details see reviews by Pain, 1996 (6), Proud and Denton, 1997 (8), Green and Noller, 1996 (9), and Sachs *et al.*, 1997 (10).

One of the contributing factors for arresting the 43 S preinitiation complex at an AUG is the sequence flanking the AUG. For vertebrates, the sequence GCCA/GCC$\underline{\text{AUG}}$G is a strong context for translational initiation, and deviations from this sequence decrease the efficiency of translation (2, 3, 13, 14, 15). The most conserved nucleotide is a purine, usually an A, at position −3. Translational initiation sites from eukaryotes and prokaryotes alike display a strong preference for a purine at this position (4, 5, 16). Very weak contexts that lack a purine at −3 cause reduced levels of protein or significant initiation of translation from downstream AUG codons with stronger contexts (13, 15). The mechanisms by which the 43 S complex interacts with translational initiation sites are poorly understood, except for the requirement for the Met-tRNA anticodon and evidence for involvement of eIF2 (6). The effects of context on the efficiency of trans-

Received 17 September 1997; accepted 31 August 1998.

* Author to whom correspondence should be addressed. E-mail: mcclint@pop.uky.edu

Crustaceans, n = 98

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	31	26	27	24	31	24	41	42	33	36	21	30	35	24	28	65	39	20		33	26	8	28	33	18	23	21	25	36
C	24	30	26	34	24	39	23	17	29	21	40	30	17	33	46	5	27	42		23	28	17	19	21	37	18	22	31	12
G	18	16	17	13	18	20	15	18	23	12	21	13	24	15	11	27	16	28		35	26	42	40	27	21	37	32	21	43
T	28	28	30	29	27	18	21	23	16	31	19	26	24	28	15	3	18	10		9	21	33	13	20	25	23	26	24	8
	a	c	c	c	a	c	a	a	a	a	c	a/c	a	c	c	A	a	c	ATG	g	c	G/T	g	a	c	g	g	c	A/G

Annelids, n = 51

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	30	28	37	54	28	37	25	51	45	38	26	44	46	20	38	75	45	24		20	18	22	32	26	36	24	30	34	26
C	15	17	26	13	26	17	19	14	20	26	32	26	12	28	36	2	16	29		16	40	22	22	28	20	30	34	10	16
G	17	26	13	17	22	17	27	16	8	10	14	18	16	12	16	22	12	45		32	18	24	22	28	24	28	16	14	40
T	37	28	24	15	24	28	29	18	27	26	28	12	26	40	10	2	28	2		32	24	32	24	18	20	18	20	42	18
	t	a/t	a	A	a	a	t	A	a	a	c	a	a	t	a	A	a	g	ATG	g/t	c	t	a	c/g	a	c	c	A/T	g

Molluscs, n = 249

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	35	37	31	31	31	38	35	39	50	34	33	39	31	19	32	73	39	35		28	34	18	25	28	23	28	29	26	35
C	20	23	27	23	24	30	29	24	19	19	34	22	14	33	45	3	30	35		18	37	14	20	26	37	18	23	25	20
G	16	15	19	20	17	7	13	17	10	27	16	12	24	16	10	18	10	20		38	18	29	37	18	18	28	13	21	28
T	29	25	23	26	20	25	22	20	21	20	17	27	31	32	12	5	21	10		17	12	39	18	28	22	26	36	29	17
	a	a	a	a	a	a	a	a	A	a	c	a	a/t	c	A/C	A	a	a/c	ATG	g	c	t	g	a/t	c	a	t	t	a

Aplysia californica, n = 66

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	11	27	24	31	20	25	28	35	42	30	33	48	33	19	21	75	30	33		20	35	14	27	30	14	21	32	29	33
C	40	25	24	22	20	35	35	28	30	18	37	21	14	30	48	0	33	38		20	41	18	15	27	35	17	21	27	24
G	37	27	29	31	27	8	13	15	13	43	18	19	27	18	10	18	8	13		46	17	38	35	14	29	38	23	18	27
T	12	20	24	17	32	32	23	22	15	10	13	13	25	33	22	8	29	16		15	8	30	23	29	23	24	24	26	15
	C/G	a/g	g	a/g	t	c	c	a	a	g	c	a	a	t	c	A	c	c	ATG	g	A/C	g	g	a	c	g	a	a	a

Echinoderms, n = 230

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	33	41	37	36	33	39	31	41	46	31	33	43	38	19	32	84	33	22		15	20	20	20	29	24	34	27	32	28
C	27	18	27	25	24	31	27	23	17	22	28	24	13	24	54	2	28	50		15	51	12	19	24	31	21	28	27	18
G	18	19	16	16	11	14	12	14	16	17	12	14	21	13	6	9	8	18		40	19	25	46	30	16	30	23	15	39
T	22	22	21	23	31	15	30	22	21	30	27	20	28	44	8	5	31	9		29	10	44	16	17	30	15	22	26	15
	a	a	a	a	a	a	a	a	a	a	a	a	a	t	A/C	A	a	C	ATG	g	C	t	g	g	c	a	c	a	g

Strongylocentrotus purpuratus, n = 91

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	33	39	34	35	31	36	30	39	54	25	29	33	31	18	38	78	37	20		20	26	23	22	26	12	30	28	32	30
C	21	16	28	22	26	35	30	21	17	24	19	29	17	28	48	4	26	48		14	44	10	14	24	31	24	24	26	18
G	24	23	21	16	14	11	13	16	12	23	14	12	25	9	7	14	8	20		39	19	27	45	26	25	27	20	18	35
T	23	23	18	28	29	19	27	24	18	29	38	26	28	45	7	5	29	13		27	11	40	19	23	32	19	28	25	17
	a	a	a	a	a	a	a/c	a	A	t	t	a	a	t	A/C	A	a	c	ATG	g	c	t	g	a/g	t	a	a/t	a	g

Cnidarians, n = 85

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6	+7	+8	+9	+10	+11	+12	+13
A	26	37	42	34	45	47	36	37	37	41	41	47	42	23	48	69	47	43		29	32	24	32	37	34	34	31	37	29
C	21	19	11	14	18	9	27	16	21	14	12	15	8	31	21	10	15	28		7	39	11	16	18	18	18	18	13	19
G	16	12	16	10	14	12	8	20	17	14	13	12	21	18	13	13	12	10		35	11	21	31	17	20	33	17	15	33
T	37	32	31	43	24	32	29	27	24	31	35	26	30	29	19	9	26	19		28	18	44	22	28	28	15	34	35	18
	t	a	a	A/T	a	A/T	a	a	a	a	A/T	a	a	c	a	A	a	a	ATG	g	c	t	a	a	a	a	a	a	g

Figure 1. Nucleotide frequencies at positions flanking translational initiation sites of marine invertebrate taxonomic groups. Uppercase letters depict nucleotides that meet the 50/75 consensus rule (18). At all other positions the nucleotide (or nucleotides in cases of equal frequencies) with the highest frequency is shown in lowercase. All nucleotides meeting the 50/75 consensus rule were found by chi-square analysis to differ significantly from expected frequencies.

lational initiation are, however, incremental rather than absolute, because even weak contexts support the initiation of translation. In fact, the vast majority of vertebrate cDNAs possess translational initiation sites that are not identical to the strong context of GCCA/GCCATGG (4, 17).

Nucleotide preferences in sequences that flank translational initiation sites vary among major taxonomic groups (4, 5). Therefore, we have analyzed the frequencies of nucleotides at positions flanking translational initiation sites of cDNAs from several major marine invertebrate taxa (phylum or subphylum) that have been diverging for many millions of years. We have found that nucleotide frequencies at these positions are phyletically distinct.

Sequence databases now contain sufficient information from several marine invertebrate phyla that tables of nucleotide frequencies at translational initiation sites can be constructed. Using the sequence databases of the National Center for Biotechnology Information, we analyzed translational initiation sites in sequences from the phyla Cnidaria (376 sequences), Echinodermata (1053), Mollusca (1365), Annelida (270), and the subphylum Crustacea (690). Because many of the sequences analyzed were fragments or were ribosomal or mitochondrial DNAs, the number of sequences with suitable translational initiation sites was much smaller: 85 for cnidarians, 230 for echinoderms, 249 for molluscs, 51 for annelids, and 95 for crustaceans. In addition, sufficient sequences from *Strongylocentrotus purpuratus* and *Aplysia californica* were available that their translational initiation sites could be compared with those of all echinoderms and molluscs, respectively. For each major taxon, the most abundant protein families in the database (and their percent occurrence) were as follows: cytoskeletal elements (11.8%) in annelids, kinases (8.2%) in cnidarians, peptide hormones (14.3%) in crustaceans, histones (22.6%) in echinoderms, and bioactive peptides (11.2%) in molluscs. We believe that the types of cDNA clones in the database were sufficiently diverse as to prevent significant bias in our analyses. For each taxon, the identified sequences were individually inspected, and the sequences surrounding each translational initiation site (positions −20 to +20) were incorporated into a spreadsheet on a personal computer. Using a computer program written in Pascal, the frequencies of occurrence of nucleotides at each position were calculated.

Figure 1 shows the nucleotide frequencies at each position from −18 to +13 surrounding the translational initiation sites from each taxon. In comparisons of a single species with its own phylum, we observed that the nucleotide frequencies at corresponding positions were similar. This comparison revealed only one position in which the frequency of a nucleotide differed more than 16 percentage points. This occurred in the comparison of *A. californica* and all molluscs at a position −18, relatively distant from the ATG codon. Clear differences in nucleotide frequencies at positions near translational initiation sites are probably apparent only between distant taxa (see refs. 4, 5).

The nucleotide frequencies at each position were used to determine consensus sequences for each taxon. These sequences are shown in a format (5) that uses uppercase letters for nucleotides that have met the criteria of the 50/75 consensus rule, and lowercase letters for the nucleotide with the highest frequency (the preferred nucleotide) in positions at which no nucleotide reached consensus (Fig. 1 and Table I). The 50/75 rule (18) specifies that a nucleotide reaches consensus if its frequency is greater than 50% and more than twice the frequency of any other nucleotide at that position. If the sum of the frequencies of any two nucleotides at one position is greater than 75%, the nucleotides are assigned co-consensus. The marine invertebrates conformed to the strong preference for a purine at position −3, with A reaching consensus at this position in all taxa (Fig. 1). Few other positions reached consensus in the marine invertebrate taxa, consistent with previous analyses of other taxonomic groups (4, 5).

These similarities are, however, overshadowed by differences among the consensus translational initiation site contexts of the five marine invertebrate taxa (Fig. 1 and Table I). These differences include the positions and identities of consensus nucleotides, and trends in preferences for specific nucleotides. The number of positions at which A is the preferred nucleotide correlates positively with early divergence in the inferred phylogenetic relationships of the marine invertebrates, yeast, and vertebrates (19). In yeast, A is the preferred nucleotide at every position from −10 to −1 (4). The cnidarians had A as the preferred

Table I

Consensus translational initiation sites

	Position						
Taxon	−6	−5	−4	−3	−2	−1	+4
Vertebrata (1)	g	c	c	A/G	c	C	A T G g
Drosophila (4)	a	c	C	A	a	a	A T G g
Crustacea	a	c	c	A	a	c	A T G g/a
Annelida	a	t	a/c	A	a	g	A T G g/t
Mollusca	a/t	c/t	A/C	A	a	a/c	A T G g
Echinodermata	a	t	A/C	A	a/t	C	A T G g
Cnidaria	a	a/c	a	A	a	a	A T G g
Yeast (4)	a	a	a	A	a	a	A T G t

The preferred nucleotide at each position is shown. In cases where the two nucleotides with the highest frequencies at a position were separated by less than two percentage points, or where the 75% rule was met, both nucleotides are shown. Nucleotides that met the 50/75 consensus rule (18) are shown in uppercase.

nucleotide at nine positions over this region; the molluscs, echinoderms, and annelids had seven; and the crustaceans (and *Drosophila*, ref. 4) had six. Vertebrates have A as the preferred nucleotide at only one of these positions (4). Interestingly, the consensus translational initiation sites of plant phyla also show a similar divergence, with a preference for A content in the dicot plants and for G/C content in the monocot plants (4, 5).

We have found that invertebrate phyla representing radial metazoans and the deuterostome and protostome groups of bilateral metazoans have different consensus sequence contexts at translational initiation sites. Consensus for A at position −3 is a common element in sequences that otherwise tend to be distinct. The consensus sequences that we have generated (Table I) are one of several factors that can aid in the identification of translational initiation sites in marine invertebrate phyla.

Acknowledgments

Supported by NIH award DC 02366 and a Lucy B. Lemann Fellowship at the Marine Biological Laboratory to T. S. Mc.

Literature Cited

1. **Kozak, M. 1983.** Translation of insulin-related polypeptides from messenger RNAs with tandemly reiterated copies of the ribosome binding site. *Cell* **34:** 971–978.
2. **Kozak, M. 1986.** Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292.
3. **Kozak, M. 1987.** At least six nucleotides preceding the AUG initiator codon enhance translation by eukaryotic ribosomes. *J. Mol. Biol.* **196:** 947–950.
4. **Cavener, D. R., and S. C. Ray. 1991.** Eukaryotic start and stop translation sites. *Nucleic Acids Res.* **19:** 3185–3192.
5. **Joshi, C. P., H. Zhuo, X. Huang, and V. L. Chang. 1997.** Context sequences of translation initiation codon in plants. *Plant Mol. Biol.* **35:** 993–1001.
6. **Pain, V. M. 1996.** Initiation of protein synthesis in eukaryotic cells. *Eur. J. Biochem.* **236:** 747–771.
7. **Sprengart, M. L., and A. G. Porter. 1997.** Functional importance of RNA interactions in selection of translation initiation codons. *Mol. Microbiol.* **24:** 19–28.
8. **Proud, C. G., and R. M. Denton. 1997.** Molecular mechanisms for the control of translation by insulin. *Biochem. J.* **328:** 329–341.
9. **Green, R., and H. F. Noller. 1997.** Ribosomes and translation. *Annu. Rev. Biochem.* **66:** 679–716.
10. **Sachs, A. B., P. Sarnow, and M. W. Hentze. 1997.** Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell* **89:** 831–838.
11. **Dever, T. E., L. Feng, R. C. Wek, A. M. Cigan, T. F. Donahue, and A. G. Hinnebusch. 1992.** Phosphorylation of initiation factor 2α by protein kinase GCN2 mediates gene-specific translational control of *GCN4* in yeast. *Cell* **68:** 585–596.
12. **Sachs, A. B., and S. Buratowski. 1997.** Common themes in translational and transcriptional regulation. *Trends Biochem. Sci.* **22:** 189–192.
13. **Kozak, M. 1989.** Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.* **9:** 5073–5080.
14. **Kozak, M. 1994.** Determinants of translational fidelity and efficiency in vertebrate mRNAs. *Biochimie* **76:** 815–821.
15. **Feng, Y., L. E. Gunter, E. L. Organ, and D. R. Cavener. 1991.** Translational initiation in *Drosophila melanogaster* is reduced by mutations upstream of the AUG initiator codon. *Mol. Cell. Biol.* **11:** 2149–2153.
16. **Kozak, M. 1987.** An analysis of 5′ noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15:** 8125–8148.
17. **Kozak, M. 1996.** Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* **7:** 563–574.
18. **Cavener, D. R. 1987.** Comparison of the consensus sequences flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* **15:** 1353–1361.
19. **Wainright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel. 1993.** Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* **260:** 340–342.